Local compressed convex spectral embedding for bird species identification

Anshul Thakur, Vinayak Abrol, Pulkit Sharma, and Padmanabhan Rajan

Citation: The Journal of the Acoustical Society of America **143**, 3819 (2018); doi: 10.1121/1.5042241
View online: https://doi.org/10.1121/1.5042241
View Table of Contents: http://asa.scitation.org/toc/jas/143/6
Published by the Acoustical Society of America

Local compressed convex spectral embedding for bird species identification

Anshul Thakur,[a] Vinayak Abrol, Pulkit Sharma, and Padmanabhan Rajan
School of Computing and Electrical Engineering, IIT Mandi, Mandi, Himachal Pradesh-175005, India

(Received 30 November 2017; revised 14 April 2018; accepted 18 April 2018; published online 29 June 2018)

This paper proposes a multi-layer alternating *sparse–dense* framework for bird species identification. The framework takes audio recordings of bird vocalizations and produces compressed convex spectral embeddings (CCSE). Temporal and frequency modulations in bird vocalizations are ensnared by concatenating frames of the spectrogram, resulting in a high dimensional and highly sparse super-frame-based representation. Random projections are then used to compress these super-frames. Class-specific archetypal analysis is employed on the compressed super-frames for acoustic modeling, obtaining the convex-sparse CCSE representation. This representation efficiently captures species-specific discriminative information. However, many bird species exhibit high intra-species variations in their vocalizations, making it hard to appropriately model the whole repertoire of vocalizations using only one dictionary of archetypes. To overcome this, each class is clustered using Gaussian mixture models (GMM), and for each cluster, one dictionary of archetypes is learned. To calculate CCSE for any compressed super-frame, one dictionary from each class is chosen using the responsibilities of individual GMM components. The CCSE obtained using this GMM-archetypal analysis framework is referred to as local CCSE. Experimental results corroborate that local CCSE either outperforms or exhibits comparable performances to existing methods including support vector machine powered by dynamic kernels and deep neural networks.
 https://doi.org/10.1121/1.5042241

[PG] Pages: 3819–3828

I. INTRODUCTION

Birds play many important roles in upholding ecological balance, from maintaining the forest cover by seed dispersal and pollination, to occupying various levels in the food chain.[1] However, due to human-induced climate change and habitat destruction, many bird species are facing the threat of population decline.[2] This has led to several conservation efforts, of which surveying and monitoring are integral components. These include maintaining records of avian diversity and the populations of various species in a particular area of interest.[3] The manual surveying of birds in their natural habitat can be difficult as birds occupy a wide range of habitats. Moreover, it is time-consuming, expensive and experienced bird watchers are required. Thus, there is a need to develop automatic methods for surveying birds in their natural habitat.

Acoustic communication in birds is very rich,[4] hence, the presence of many birds species can be detected by analyzing their sounds or vocalizations. This makes acoustic monitoring a convenient and passive method to monitor birds in their respective habitats. Recent advancements in programmable recording devices have made acoustic monitoring feasible. These devices can record a large amount of acoustic data, which can be used for monitoring avian diversity. In this work, we target the problem of bird species identification from recorded acoustic data, which forms the backbone of an acoustic monitoring system.

Various methods have been proposed in the literature for the problem of bird species identification/classification from recorded bird songs or calls. In an initial study, McIlraith and Card[5] proposed to use a two-layer feed forward neural network with back propagation for bird song classification. Harma and Somervuo[6–8] used sinusoidal modeling of syllables (the basic unit of bird song) for species classification. Fagerlund[9] proposed a decision tree-based hierarchical classification framework for bird species recognition, where each node of the tree is a support vector machine (SVM). The feature representation used is Mel frequency cepstral coefficients (MFCC) and low-level signal descriptors. Lee *et al.*[3] proposed to use two-dimensional cepstral coefficients for bird species identification. Their study also proposed to tackle within-class variation by prototyping each class using vector quantization and Gaussian mixture models. Stowel and Plumbley[10] proposed a spherical K-means-based unsupervised representations for bird species classification. Apart from these methods, many studies have targeted various bioacoustic problems using deep learning, e.g., deep convolution neural networks (CNN) have been used for bird species identification.[11–14] Chakraborty *et al.*[15] utilized a three-layered deep neural network (DNN) for bird species classification, where MFCCs are used as the feature representation. Apart from DNN, their study also explored Gaussian mixture model (GMM), GMM-UBM (universal background model) and SVM powered by various dynamic kernels[16] for species identification.

Leveraging on the success of learned-feature representations obtained by factorizing spectrograms for acoustic scene classification[17] and acoustic event detection,[18] we propose a

[a]Electronic mail: anshul_thakur@students.iitmandi.ac.in

supervised, multi-layer, alternating dense-sparse framework to obtain feature representations for bird species identification. In the proposed method, a given recorded audio signal (*dense*) is converted into a magnitude spectrogram (*sparse*). This concept of sparsity comes from the analysis that most of the bird vocalizations usually occupy only a few frequency bins in the spectrogram.[19] The frequency and temporal modulations present in bird vocalizations provide species-specific signatures. However, applying matrix factorization techniques on spectrograms directly may not capture these modulations effectively. To overcome this issue, a certain number of frames are concatenated around each frame of the spectrogram for embedding the context. This results in a high dimensional (*sparse*) super-frame representation that is capable of capturing the frequency and temporal modulations more effectively. These high dimensional super-frames are unsuitable for acoustic modeling due to high computational complexity. Since the spectrogram is sparse, this super-frame representation is also sparse. Hence, super-frames can be compressed without losing too much information. Random projections,[20] which preserve pairwise distance according to the Johnson–Lindenstrauss (J–L) lemma, are used to compress these super-frames to a low-dimensional representation (*dense*). In the next step, the vocalizations of each bird species are modeled using restricted robust archetypal analysis (AA). AA provides compact, probabilistic and interpretable representation[23] in comparison to the other matrix factorization techniques such as non-negative matrix factorization (NMF) and sparse dictionary learning.[22] The learned archetypal dictionaries are used to obtain a sparse-convex representation for the compressed super-frames. These representations are designated as compressed convex spectral embeddings (CCSE). This CCSE representation captures species-specific signatures effectively and can be used as feature representation in any classification framework.

CCSE assumes that the compressed super-frames of a bird species lie on only one manifold. However, a particular bird species can have a large repertoire of vocalizations that often occupy different manifolds in the feature space.[3] Therefore, a single archetypal dictionary per bird species may not be able to model the variations present in a single bird class. We address this problem by proposing to use multiple archetypal dictionaries to model one bird species. In order to learn multiple dictionaries, the compressed super-frames are clustered using GMM and for each cluster, an individual archetypal dictionary is learned. To obtain the CCSE for a compressed super-frame, a dictionary is chosen for each class using the responsibility terms of the class-specific GMM. The CCSE obtained using this GMM-AA-based framework is designated as local CCSE.

The archetypes learned using AA approximates the convex-hull of the data, and the estimation of these archetypes is often expensive in terms of computation.[24] Hence, in order to speed up the process of finding archetypes, we use a restricted version of AA. In restricted AA, only the data points around the convex hull/boundary are used for determining the archetypes. Conventionally, AA is performed individually for each class and without any separate effort to increase the inter-class discrimination. Hence, there can be a high correlation between atoms of inter-class dictionaries, which may degrade the discriminative ability of these dictionaries. Supervised dictionary learning methods such as label-consistent K-singular value decomposition (Ref. 25) overcome this problem by learning dictionaries in a supervised manner. Nevertheless, these supervised dictionary learning techniques are computationally expensive (both in time and space) and are not feasible when a substantial number of classes are involved. In order to overcome this issue, we propose an efficient greedy procedure to choose atoms from each dictionary such that the overall correlation among all dictionaries is decreased. This procedure not only reduces the gross-correlation among dictionaries but also helps in reducing their size. Decreasing the dictionary size reduces the computational complexity, which can be helpful for large-scale species identification.

The major contributions of this work are summarized as follows:

(1) Local CCSE, a supervised feature representation, that handles intra-class variations efficiently (Algorithm 2).
(2) The application of a restricted version of archetype analysis for acoustic modeling.
(3) A greedy procedure to choose a subset of atoms from each dictionary such that the overall correlation among all local dictionaries of all classes is reduced (Algorithm 1).

The rest of this paper is organized as follows. In Sec. II, we describe CCSE-based framework. In Sec. III, the proposed local CCSE framework along with the proposed pruning procedure to decrease the inter dictionary correlation is discussed. Experimental setup and observations are in Secs. IV and V, respectively. Section VI concludes the paper.

II. COMPRESSED CONVEX SPECTRAL EMBEDDINGS (CCSE)

In this section, the overall process to obtain CCSE from any input recording is described (Fig. 1). First, we describe the process of obtaining a compressed super-frame-based representation from any input audio recording. Then, we explain the procedure to learn an archetypal dictionary for each bird species. Finally, we describe the process to obtain CCSE for any audio recording.

A. Computing compressed super-frames

The short time Fourier transform (STFT) is applied to obtain a magnitude spectrogram $\mathbf{S}\,(m \times N, m$ is the number of frequency bins, N is the number of frames) from each input audio recording. Short term Fourier analysis often leads to the smearing of temporal and frequency modulations present in bird vocalizations. In order to capture these modulations more effectively, context information is ingrained into the current frame (under processing) of the spectrogram by concatenating W previous and W next frames around the current frame. This concatenation produces a high dimensional $[(2Wm + m) \times 1]$ representation called a super-frame. The pooled spectrograms of all the training examples of a particular class, $\hat{\mathbf{S}}\,(m \times l, m$ is the number of frequency bins, l is the number of pooled frames), are converted into super-frame representation, $\mathbf{F} \in \mathbb{R}^{(2Wm+m)\times l}$, using the aforementioned concatenation

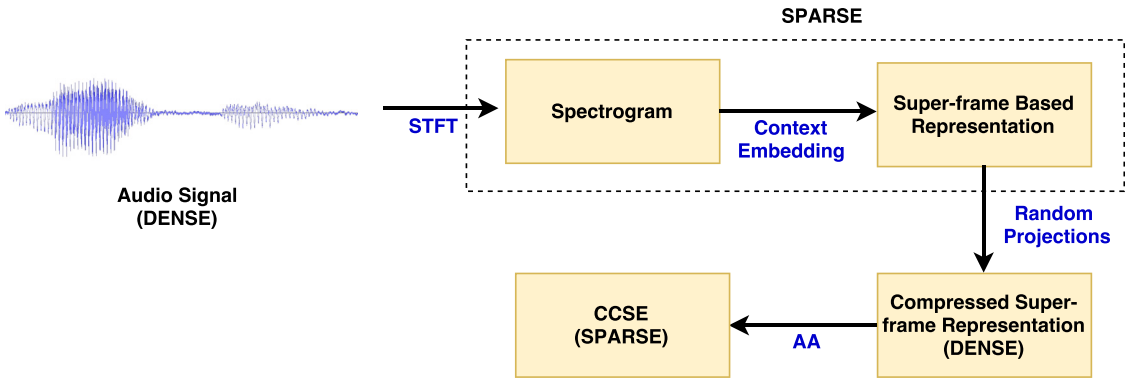

FIG. 1. (Color online) Proposed pipeline for obtaining CCSE from an audio signal.

process. These super-frames are high-dimensional, which makes them computationally expensive to process for acoustic modeling. However, these super-frame representations are sparse. The sparsity of the spectrogram and super-frames is illustrated in Fig. 2. Due to this sparsity, super-frames are suitable to attain a high degree of compression. Hence, building upon the J–L lemma,[21] random projections are used to compress these super-frames. Gaussian random matrices satisfy the J-L lemma with high probability.[26] Hence, these random matrices preserve the pair-wise distance between super-frames in the projected space. In particular, a random Gaussian matrix $\mathbf{G}$ (of dimensions $K \times 2Wm + m$) is used to achieve the transformation, $\phi : \mathbb{R}^{2Wm+m} \rightarrow \mathbb{R}^{K}$, which compresses the super-frames. This compressed representation, $\mathbf{X} = \mathbf{G} \times \mathbf{F}$, $\mathbf{X} \in \mathbb{R}^{K \times l}$, is used to learn the archetypal dictionaries. Figure 2(c) depicts the compressed super-frame representation obtained for the spectrogram shown in Fig. 2(a). Similarly, for a test audio recording, compressed super-frames are obtained using the same procedure.

B. Restricted robust AA for dictionary learning

The CCSE framework employs archetypal analysis (AA) for acoustic modeling. The compressed super-frames corresponding to the bird vocalization regions are used for learning the archetypes. The bird vocalization regions are identified (in the input recordings) using a semi-supervised segmentation method[27] proposed in one of our earlier studies. Using AA, which is a non-negative matrix factorization technique, the matrix of compressed super-frames, $\mathbf{X}$, is decomposed to obtain the representation matrix $\mathbf{A}$ as $\mathbf{X} = \mathbf{DA}$. The dictionary, $\mathbf{D}$, consists of the archetypes, which lie on the convex hull of data. These archetypes are confined to be the convex combination of the individual data points, i.e., $\mathbf{D} = \mathbf{XB}$, $\mathbf{D} \in \mathbb{R}^{K \times d}$ (d is the number of archetypes) and $\mathbf{B} \in \mathbb{R}^{l \times d}$.

1. Restricting AA

Generally, matrix factorization is a computationally expensive process and AA is no exception. However, it is

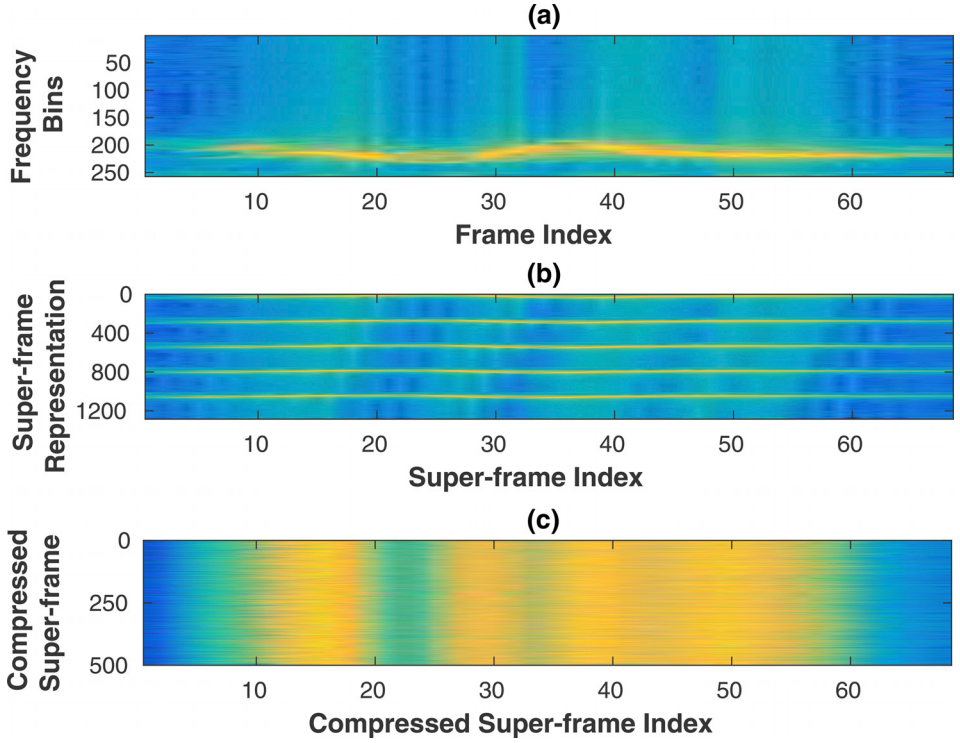

FIG. 2. (Color online) (a) Spectrogram of a Cassin's vireo vocalization, (b) 1285-dimensional ($2Wm + m = 1285$ and $m = 257$) super-frame representation obtained from (a) using $W = 2$. W is window size for concatenation and m is the number of frequency bins. (c) Compressed super-frames of 500 dimensions (K = 500) obtained by projecting (b) on a random Gaussian matrix.

known that archetypes lie on the boundary or convex hull of the data. This property can be used to restrict the archetypal search space to the data points existing around the boundary. This restricted search reduces the computational time required to learn the archetypes.

Let $\mathcal{B}$ be the index of compressed super-frames that lie around the boundary. To find these super-frames, the following objective function is minimized:

$$||\mathbf{X} - \mathbf{XC}||_F^2 \text{ s.t. } \text{diag}(\mathbf{C}) = 0,$$
$$\mathbf{c}_i \geqslant 0, \quad \text{and} \quad ||\mathbf{c}_i||_1 = 1, \tag{1}$$

where diag(.) denotes the diagonal elements. The solution $\mathbf{C}$ (having columns $\mathbf{c}_i$) that minimizes the given objective function, can be interpreted as the coefficient matrix for representing each compressed super-frame ($\mathbf{x}_i$) in $\mathbf{X}$ as a linear combination of other compressed super-frames.[24] The significant values (i.e., high magnitude values) of the solution correspond to the boundary points $\mathbf{x}_z$, such that $z \in \mathcal{B}$. These values are obtained by maximizing the negative gradient of the error cost in Eq. (1) (involving inner products) with respect to $\mathbf{c}_i$. The principles of convex geometry state that the inner product between two points is maximum when one of the points lies on boundary of the data.[28] As a result, the solution that minimizes the error cost in Eq. (1) ensures that the union of the indices of high magnitude elements of each $\mathbf{c}_i$ refer to super-frames around the boundary. Hence, using this procedure $\mathbf{X} \in \mathbb{R}^{K \times l}$ is reduced to $\hat{\mathbf{X}} \in \mathbb{R}^{K \times p}$ (p is the number of chosen boundary super-frames such that $p \ll l$). The problem in Eq. (1) can be solved using a fast quadratic programming (QP) solver such as MATLAB's *quadprog* and is a one-time procedure.

2. Restricted robust AA

The presence of outliers in data changes the convex hull, which affects the performance of AA. The outliers can arise due to noise or segmentation errors. In order to address this issue, we propose to use robust AA (Ref. 23) on $\hat{\mathbf{X}}$, which mitigates the affects of outliers to a large extent. In particular, the archetypal dictionary, $\mathbf{D}$, is computed by optimizing the following function:[23]

$$\underset{\substack{\mathbf{B},\mathbf{A} \\ \mathbf{b}_j \in \Delta_p, \mathbf{a}_i \in \Delta_d}}{\text{argmin}} \sum_{i=1}^{p} h\left(||\mathbf{x}_i - \mathbf{D}\mathbf{a}_i||_2\right)$$
$$= \frac{1}{2}\sum_{i=1}^{p} \frac{1}{w_i}||\mathbf{x}_i - \hat{X}B\mathbf{a}_i||_2^2 + w_i,$$
$$\Delta_p \triangleq \left[\mathbf{b}_j \geqslant 0, ||\mathbf{b}_j||_1 = 1\right], \quad \Delta_d \triangleq \left[\mathbf{a}_i \geqslant 0, ||\mathbf{a}_i||_1 = 1\right],$$
$$w_i \geq \epsilon$$
$$\forall i : 1 \rightarrow p \quad \text{and} \quad \forall j : 1 \rightarrow d. \tag{2}$$

Here $\mathbf{x}_i, \mathbf{a}_i$, and $\mathbf{b}_j$ are the columns of $\hat{\mathbf{X}} \in \mathbb{R}^{k \times p}$, $\mathbf{A} \in \mathbb{R}^{d \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times d}$, respectively, w_i is a scalar, and ϵ is a positive constant. In contrast to conventional AA employing Euclidean loss, robust AA employs a Huber loss function $h()$. For scalars u and ϵ, the Huber function is defined as $h(u) = 1/2\min_{w \geq \epsilon}[u^2/w + w]$.[23] The use of Huber loss introduces a weight $w_i = \max(||\mathbf{x}_i - \hat{\mathbf{X}}B\mathbf{a}_i||_2, \epsilon)$ for $\mathbf{x}_i$ in the optimization process, i.e., w_i weighs the contribution of $\mathbf{x}_i$ in the estimation of archetypes. After the optimization, the weight w_i becomes larger for the outliers, reducing their importance in finding the archetypes. In this work, the optimization problem in Eq. (2) is solved using an iterative procedure proposed by Chen *et al.*[23] (algorithm 3 in Ref. 23).

3. Computational efficiency

The computational saving obtained using restricted AA is highlighted in Fig. 3. The average running times recorded for learning 32-archetypes from different number of super-frames, using restricted robust AA and traditional robust AA, are depicted in Fig. 3. This experiment is conducted on a PC running Ubuntu 16.0 with 16 Gb of RAM, and an Intel i7 CPU with 3.00 GHz clock speed. The implementation is in Matlab 2014a. Each super-frame is of 500-dimensions and 100 iterations are used for learning the archetypes for both the setups. The analysis of Fig. 3 shows that for all configurations, the average running time for restricted robust AA is significantly less than the robust AA. The restricted AA shows a relative drop of 67.5% in average running time across all configurations.

C. Computing CCSE representation

The compressed super-frames are obtained for an audio recording using the procedure discussed in Sec. II A. Here, the vocalization regions are identified and super-frames corresponding to these regions are extracted. The same Gaussian random matrix is employed for obtaining compressed super-frames during training and testing. The final dictionary, $\mathbf{D}$, is obtained by concatenating individual dictionaries of each bird species/class, i.e., $\mathbf{D} = [\mathbf{D}^1\mathbf{D}^2 \cdots \mathbf{D}^q]$, where $\mathbf{D}^q$ is the archetypal dictionary learned for the qth class using restricted robust AA (discussed in Sec. II B 3). The CCSE for any compressed super-frame, $\mathbf{y}_i$, is obtained by projecting $\mathbf{y}_i$ on to a simplex corresponding to dictionary $\mathbf{D}$, as further described in Sec. III. This CCSE contain strong

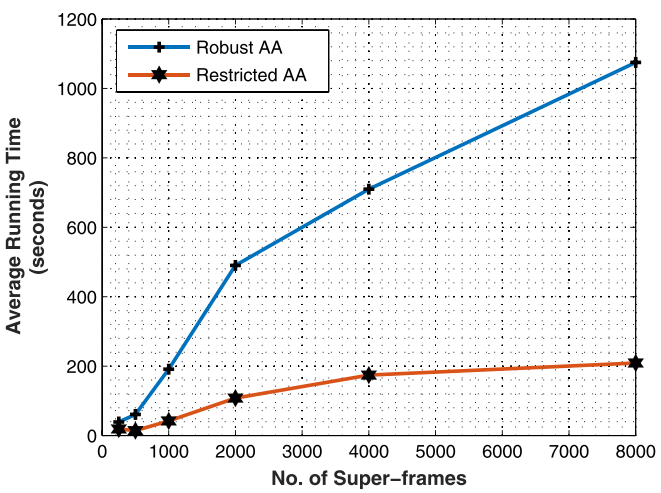

FIG. 3. (Color online) Average running time recorded for robust AA and restricted AA.

class-specific signatures and can be used as a feature representation for species classification. This behavior is illustrated in Fig. 4, which shows the average of CCSEs obtained for an exemplar vocalization of three different species. These average CCSEs are obtained using the final dictionary (**D**) derived from the individual dictionaries of all three species. The final dictionary contains 128 atoms per class (the first 128 for black-throated tit, the next 128 for black-yellow grosbeak and the last 128 for black-crested tit). In average CCSEs, the coefficients exhibit higher amplitude for the atoms of **D** which correspond to the true class. This corroborates our claim of the discriminative nature of CCSE.

III. PROPOSED LOCAL CCSE-BASED FRAMEWORK

Songs phrases and various calls such as alarm calls, feeding calls and flight calls form the repertoire of vocalizations that a species can produce. The nature of different kind of vocalizations can vary considerably.[3] A single archetypal dictionary (as used in CCSE) cannot effectively model all these within-class variations. An effective way to handle this problem is to learn local archetypal dictionaries. The CCSE learned from these local dictionaries provide better representation for a bird species. Keeping these facts in account and improvising over the CCSE framework, we propose a local CCSE-based framework which can handle the variations present in vocalizations of various bird species. In this framework, multiple local dictionaries are learned for each class. The different local dictionaries model the different sets of vocalizations of a particular species. Out of these local dictionaries, one dictionary per class is chosen to obtain convex sparse representations (CCSE) for a super-frame. This framework also utilizes a greedy iterative procedure to decrease the gross correlation between intra and inter-dictionary atoms. This reduces the size of dictionaries making the proposed framework computationally efficient.

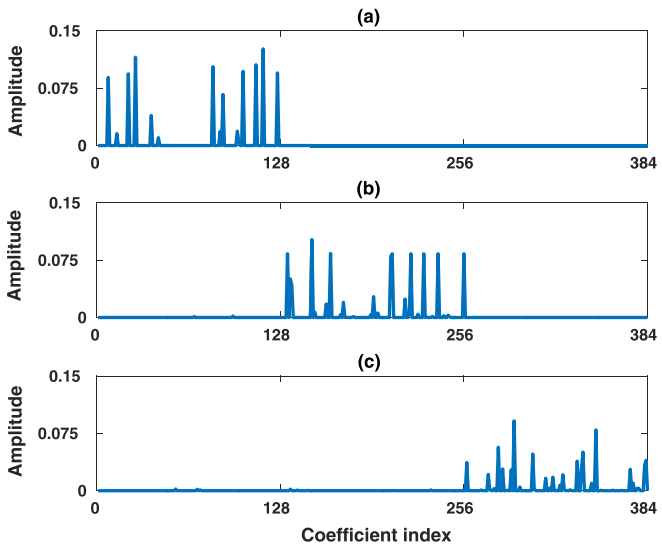

FIG. 4. (Color online) Average CCSEs obtained for a vocalization of (a) black-throated tit, (b) black-yellow grosbeak and (c) black-crested tit. Each bird species is modeled by an archetypal dictionary having 128 atoms.

A. Learning local dictionaries

The compressed super-frames corresponding to the bird vocalizations present in the training audio recordings are extracted and pooled together in a class-specific manner as described in Sec. II A. These pooled super-frames are used for learning multiple local dictionaries of a bird class. First, a GMM with Z components is used to cluster these super-frames. Then, restricted robust AA (Sec. II) is applied to get an archetypal dictionary for each of these Z clusters. Hence, one bird species/class is modeled by Z archetypal dictionaries. It has to be noted the number of GMM components can be different for different classes, e.g., Z can be large for a class having large variations in vocalizations (e.g., Cassin's vireo) as compared to the one with less variations (e.g., Hutton vireo). Since the clusters within a class can exhibit more overlap, GMM provides better clustering than the hard-clustering techniques like K-means or K-medoids.

B. Decreasing the inter-dictionary correlation

In Sec. III A, all dictionaries are learned independently, which may lead to high correlation between the inter-dictionary atoms. This high correlation is not a big issue for the dictionaries of one class. However, if correlation is high among the dictionaries of different classes, it can affect the classification performance. In order to address this problem, a greedy pruning procedure is proposed to choose a subset of atoms from each dictionary, such that the gross correlation among all the dictionaries is decreased.

Let us denote the jth pruned dictionary of the qth class by $\mathbf{D}_j^{*q}$. The proposed algorithm starts by choosing the independent atoms from the first dictionary of the first class, $\mathbf{D}_1^1$, iteratively using the following metric:

$$i = \max_{i \notin \mathcal{Z}} ||\mathbf{d}_{1i}^1 - \mathbf{D}_{1\mathcal{Z}}^1 \mathbf{D}_{1\mathcal{Z}}^{1\dagger} \mathbf{d}_{1i}^1||_2^2 \text{ s.t. } \mathbf{D}_{1\mathcal{Z}}^{1^{\mathrm{T}}} \mathbf{D}_{1\mathcal{Z}}^1 \text{ is invertible.} \tag{3}$$

Here $\mathbf{d}_{1i}^1$ is an atom of $\mathbf{D}_1^1$, $\dagger$ denotes the pseudo-inverse, $\mathcal{Z}$ denotes the set of indices of the selected atoms and $\mathbf{D}_{1\mathcal{Z}}^1 \subset \mathbf{D}_1^1$, denotes the current set of selected atoms. Equation (3) computes the distance of an atom $\mathbf{d}_{1i}^1$ to the space spanned by the atoms in $\mathbf{D}_{1\mathcal{Z}}^1$, and selects the one which lies at maximum distance from the span of $\mathbf{D}_{1\mathcal{Z}}^1$. This atom exhibits minimum correlation to atoms present in the already selected set, $\mathbf{D}_{1\mathcal{Z}}^1$. In order to choose J atoms from $\mathbf{D}_1^1$, Eq. (3) is iterated J times. Hence, a pruned dictionary, $\mathbf{D}_1^{*^1} \subset \mathbf{D}_1^1$, is obtained. This whole procedure is repeated for each local dictionary of each class to find the uncorrelated atoms with respect to the previously selected atoms from all the dictionaries. Algorithm 1 describes the procedure to obtain the pruned versions of all the dictionaries. All local dictionaries of each class are given as input to algorithm 1. The output is a set of pruned dictionaries, each having J ($J < d$) atoms. Hence, along with correlation, this procedure also decreases the size of dictionaries, thus reducing the computational complexity of the whole framework.

ALGORITHM 1: Proposed greedy procedure to decrease the inter-dictionary correlation.

input: $\mathbf{D}_z^q, z^{th}$ dictionary of qth class
$\forall q : 1 \rightarrow Q$ (number of classes)
$\forall z : 1 \rightarrow Z_q$ (number of local dictionaries in qth class)
$\mathbf{d}_{zi}^q : i$th atom of $\mathbf{D}_z^q$
$\mathbf{J}$, the number of atoms to be selected per dictionary
$\mathcal{W} = [\,]$, set of currently selected dictionary atoms
output: $\mathbf{D}^* = [\mathbf{D}_1^{*^1}\mathbf{D}_2^{*^1}\cdots\mathbf{D}_Z^{*^1}\cdots\mathbf{D}_{Z-1}^{*^q}\mathbf{D}_Z^{*^q}]$, Set of pruned dictionaries

$\mathbf{D}^* = [\,]$, $\mathcal{W} = [\mathcal{W}\ \mathbf{d}_{11}^1]$
for $q \leftarrow 1$ **to** Q **do**
  for $z \leftarrow 1$ **to** Z_q **do**
    $\mathcal{S} = [\,]$ // Set to store indices of selected atoms
    for $j \leftarrow 1$ **to** J **do**
      $i = \arg\max_i ||\mathbf{d}_{zi}^q - \mathcal{W}\mathcal{W}^\dagger \mathbf{d}_{zi}^q||_2^2$ s.t. $\mathcal{W}^T\mathcal{W}$ is invertible
      // $\forall i : 1 \rightarrow d$ (number of atoms)
      $\mathcal{W} = [\mathcal{W}\ \mathbf{d}_{zi}^q]$
      $\mathcal{S} = \mathcal{S} \cup i$
    end
    $\mathbf{D}_z^{*q} = \mathbf{D}_z^q[:, \mathcal{S}]$
    $\mathbf{D}^* = [\mathbf{D}^*\ \mathbf{D}_z^{*^q}]$
  end
end

ALGORITHM 2: Procedure to obtain average local CCSE for a bird vocalization.

input: $\mathbf{D}_z^{*q}, \forall q : 1 \rightarrow Q, \forall z : 1 \rightarrow Z_q$
G_q, GMM of qth class, $\forall q : 1 \rightarrow Q$ (number of classes)
$\mathbf{Y}$, $(K \times I)$, compressed super-frames of a bird vocalization
output: LC_{avg}, average local CCSE for $\mathbf{Y}$ of dimensions $Qd \times 1$

for $z \leftarrow 1$ **to** I **do**
  $\mathbf{D}_f^i = [\,]$
  for $q \leftarrow 1$ **to** Q **do**
    $z = \arg\max_z \gamma_z^q(\mathbf{y}_i)$, $\forall z : 1 \rightarrow Z_q$ // Using Eq. (4)
    $\mathbf{D}_f^i = [\mathbf{D}_f^i\ \mathbf{D}_z^{*q}]$
  end
  a^i = simplex Projection $(\mathbf{D}_f^i, \mathbf{y}_i)$// Achieving convex decomposition using Active − set QP solver and y_i is ith column of $\mathbf{Y}$
end
$LC_{\text{avg}} = \frac{1}{I}\sum_{i=1}^{I} \mathbf{a}_i$

C. Computing local CCSE representation

In order to obtain the local CCSE for any super-frame $\mathbf{y}_i$, one dictionary from Z_q local dictionaries of the q_{th} class is chosen. The responsibility of each GMM component/cluster in defining $\mathbf{y}_i$ is calculated and the dictionary corresponding to the component exhibiting maximum responsibility is chosen. This is achieved using the following equation:

$$z = \underset{z}{\text{argmax}}\ \gamma_z^q(\mathbf{y}_i) = \frac{w_z^q \mathcal{N}\left(\mathbf{y}_i | \boldsymbol{\mu}_z^q, \Sigma_z^q\right)}{\sum_{p=1}^{Z} w_p^q \mathcal{N}\left(\mathbf{y}_i | \boldsymbol{\mu}_p^q, \Sigma_p^q\right)}. \tag{4}$$

Here $w_z^q, \boldsymbol{\mu}_z^q$, and Σ_z^q are the weight, the mean and the covariance of the zth GMM component of the qth class. The pruned dictionary corresponding to this zth component/cluster, i.e., $\mathbf{D}_z^{*q}$, is chosen. This procedure is iterated to select Q dictionaries, one for each class, which are used for obtaining the local CCSE. These dictionaries are concatenated to form the final dictionary $\mathbf{D}_f^i$. The local CCSE for $\mathbf{y}_i$ is obtained by projecting it on a simplex corresponding to dictionary $\mathbf{D}_f^i$, using the quadratic programming-based active-set method proposed by Chen *et al.*[23] (algorithm 2 in Ref. 23). This local CCSE exhibits high coefficient values corresponding to true class atoms of $\mathbf{D}_f^i$ and low coefficient values corresponding to the atoms of other classes (plots similar to Fig. 4 are obtained). The distinction in local CCSE for super-frames of different classes makes them an appropriate feature representation for classification.

A segmented bird vocalization is represented by average of local CCSE of all the super-frames corresponding to this vocalization. Algorithm 2 describes the procedure to obtain average local CCSE for a bird vocalization. These average local CCSEs are used as a feature representation for bird species identification. As an illustration, Fig. 5 shows the two-dimensional (2-D) plot of average local CCSEs for vocalizations of seven different bird species computed using t-distributed stochastic neighbor embedding (t-SNE).[29] It must be noted that the parameters used for obtaining these average local CCSE are for illustration purpose only and may not be optimal. In this illustration, the super-frame representation of 1285 dimensions (for $W = 2$ and NFFT $= 512$) is used. Random projections are used to obtain compressed 500-dimensional representation of these super-frames. Each species is modeled by a three-component GMM and a 32-atom dictionary is learned for each component/cluster. One such 32-atom dictionary is illustrated in Fig. 6. Hence, each vocalization is represented by 224 (32×7)-dimensional average local CCSE. The analysis of Fig. 5 makes it clear that the proposed feature representation, i.e., average local CCSE shows different characteristics for different bird species, making them suitable for bird species identification. The small overlap observed between vocalizations of grey bush chat, black-crested tit and golden bush-robin could be due to the similarity between the properties (frequency range and modulations) of the vocalizations of these species.

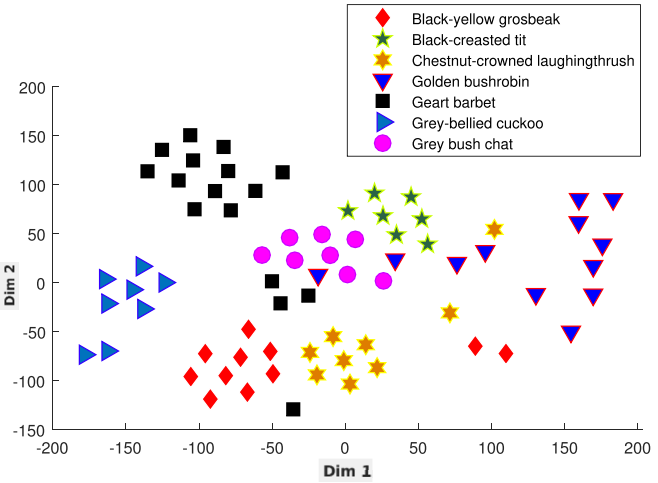

FIG. 5. (Color online) Two-dimensional t-SNE visualization of 224-dimensional average local CCSE obtained for seven different bird species.

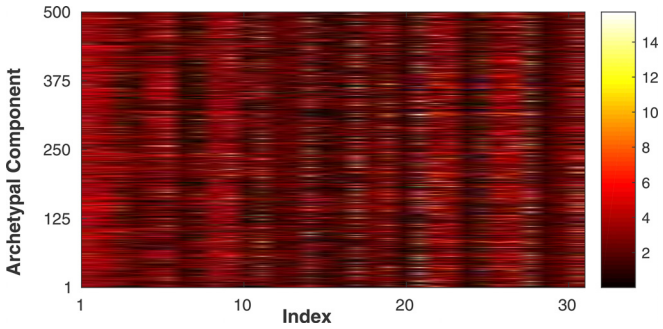

FIG. 6. (Color online) A 32-atom archetypal dictionary learned for one cluster of black-yellow grosbeak.

IV. EXPERIMENTAL SETUP

In this section, we discuss the dataset used, along with various parameters used in the experimental evaluation. In addition, the methods used for comparative study are also listed here.

A. Dataset used

Audio recordings containing vocalizations of 50 different bird species are used for evaluating the classification performance of the proposed local CCSE. These audio recordings are obtained from three different sources. Recordings of 26 bird species were obtained from the Great Himalayan national park (GHNP), in north India. These recordings were collected manually using a directional microphone. The recordings of seven bird species were obtained from the bird audio database maintained by the Art & Science center, UCLA.[30] The audio recordings of the remaining 17 bird species were obtained from the Macaulay Library.[31] These recordings are provided on an academic research license. All the recordings available are 16-bit WAV files having a sampling rate of 44.1 kHz, with the duration ranging from 18 s to 3 min. Although most of the recordings are mono channel, dual channel recordings are also present, of which the first channel is used here. The information about these 50 species along with the total number of recordings and vocalizations per species is available at http://goo.gl/cAu4Q1.

B. Parameter setting

In our experiments, each recording is converted to spectrogram using STFT (with 512 FFT points) on a frame-by-frame basis, with a frame size of 20 ms and 50% overlap. The super-frames are obtained using a window length of seven ($W = 7$), which are compressed using random projections to have a dimension of $K = 1000$. These optimal values of window length and the dimensions of compressed super-frames are determined experimentally as discussed in Sec. V. The number of GMM components (Z_q) range from 3 to 8 for different classes. The optimal number of GMM components are selected using the Bayesian information criterion (BIC); the GMM giving least BIC is used. The number of atoms in each archetypal dictionary (learned for each GMM component) is $d = 128$. These atoms are pruned down to $J = 32$, using the procedure described in Algorithm 1. These optimal values of d and J are determined empirically. The classifier used in this work is linear SVM, with an empirically tuned penalty parameter. The average local CCSE obtained from each segmented vocalization is used as the feature representation. Hence, the proposed framework provides segment/vocalization level classification decisions.

1. Train/test data distribution

A threefold cross-validation is used to compare the classification performance of the proposed local CCSE framework and the comparative methods. 33.33% of the vocalizations present in each fold (per class) are used for training while the remaining are used for testing. 75% of these 33.33% training vocalizations are used for learning dictionaries while remaining 25% vocalizations are used to obtain the average local CCSE for training the SVM. The results presented here are averaged across all three folds.

2. Comparative methods

The classification performance of the proposed local CCSE framework is compared with GMM, GMM-UBM, SVM powered by dynamic kernels and DNN-based classifiers. Different dynamic kernels used in this study are: probabilistic sequence kernel (PSK), Gaussian mixture models super-vector kernel (GMMSV), GMM-UBM mean interval kernel (GUMI), GMM-based pyramid match kernel (PMK), and GMM-based intermediate matching kernel (IMK). The DNN used for comparison is a three layered fully connected network with 512 hidden units.[15] To tackle over-fitting, a drop-out rate of 10% is used. MFCC using delta and acceleration coefficients, with a temporal context of seven previous and seven next MFCC frames are used as feature representations in the above mentioned methods. For the GMM-based classifier, the optimal number of GMM components per class is learned using BIC. Further, a UBM built by pooling the frames of all classes and fitting a 128-component GMM, is used for the GMM-UBM method. In addition, spherical K-means-based unsupervised feature representation[10] is also used for comparison. Here, features are obtained using 500 clusters means and a random forest classifier (with 200 decision trees) is used for classification.

The performance of local CCSE is also compared with CCSE (see Sec. II). For classification, each vocalization is represented as the average of CCSE obtained for all the super-frames of that vocalization. Each class is modeled by a single dictionary having 128 archetypes and a linear SVM is used for classification purposes.

V. EXPERIMENTAL OBSERVATIONS

In this section, first, we describe the effects of size of context window, extent of compression in super-frames and size of pruned dictionaries on the classification performance of the proposed framework. Then, the classification performance of the proposed framework is evaluated against the performances of the various existing methods. Finally, the performance of the proposed framework and local CCSE is

evaluated when there is a significant mismatch in training-testing conditions.

A. Effect of context window size (*W*)

A smaller value of W leads to a super-frame representation having less context information and lower dimensionality. On the other hand, larger value of W produces super-frames having more context and high dimensionality. Although these high dimensional super-frames are compressed using random projections, obtaining a larger compression ratio may lead to the loss of information. Hence, an appropriate value of W is chosen empirically. The minimum value of W which gives the maximum classification performance can be considered as optimal. Figure 7 shows the classification performance achieved by the local CCSE-based framework for different values of W. It is clear from the figure that the incorporation of context information improves the classification performance. The maximum accuracy is achieved for $W = 7$. On increasing W further does not lead to better classification. Hence, $W = 7$ is chosen for all the experiments in this study. It must be noted that for all the values of W, a compression ratio of 75% was maintained for obtaining the compressed super-frames. Using a very large value of W ($W > 10$) can lead to over-fitting by affecting the generative nature of the proposed method, as shown in Fig. 7.

B. Compression vs classification/computation trade-off

The computational complexity of robust AA and active-set simplex decomposition is directly dependent on the dimensionality of data points.[23] Hence, reducing the dimensionality of super-frames makes the proposed framework computationally more efficient. As discussed earlier, a window size of $W = 7$ is used in our experimentation. This gives rise to 3855-dimensional super-frames [FFT points = 512, $3855 = 257 \times (7 + 1 + 7)$]. To determine the extent of compression that can be achieved in the super-frames, we experimented with different compression rates and the results are shown in Fig. 8. It can be observed that one can achieve a 75% compression ($K = 1000$ from original dimension of 3855), without any decrease in the classification accuracy. This high compression can be attributed to the highly sparse nature of the super-frames. Figure 8 also shows the increment in average running time (average time recorded for 10 runs) for learning local dictionaries of 50 classes (used in experimentation) as the dimensionality of compressed super-frames is increased. Hence, compressing the super-frames provide significant computational gain in the proposed framework.

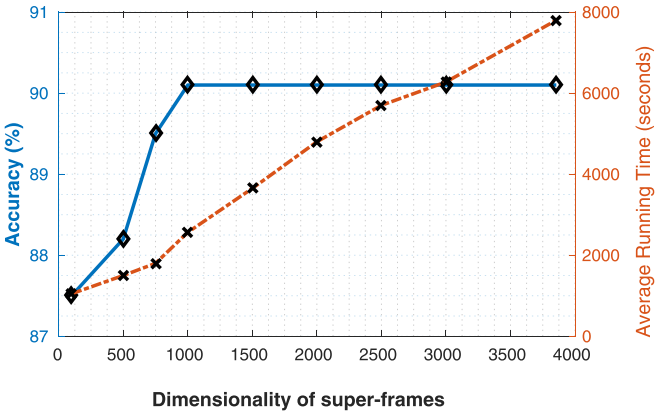

FIG. 8. (Color online) Effect of compression on classification performance and average running time required for learning local dictionaries.

C. Size of pruned dictionaries vs classification performance

The pruning procedure given in algorithm 2 decreases the size of dictionaries by choosing a subset of atoms from each dictionary. In this experiment, we analyzed the extent to which the size of dictionaries can be reduced without showing performance degradation. Originally, each dictionary has 128 atoms. We pruned these dictionaries to have 64, 32, 16, and 8 atoms. Figure 9 depicts the classification performance of local CCSE for each of these cases. It can be observed from Fig. 9 that using pruned dictionaries having 32 atoms each, provide the same classification performance as the original dictionaries.

D. Classification performance

The comparison of classification performance of the proposed local CCSE-based framework with various comparative

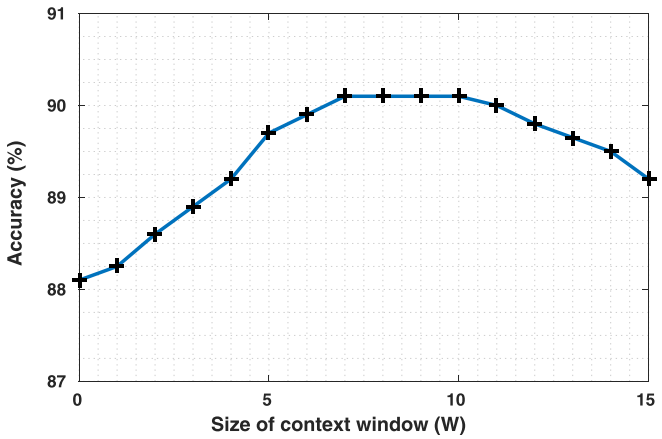

FIG. 7. (Color online) Effect of the size of context window on classification performance.

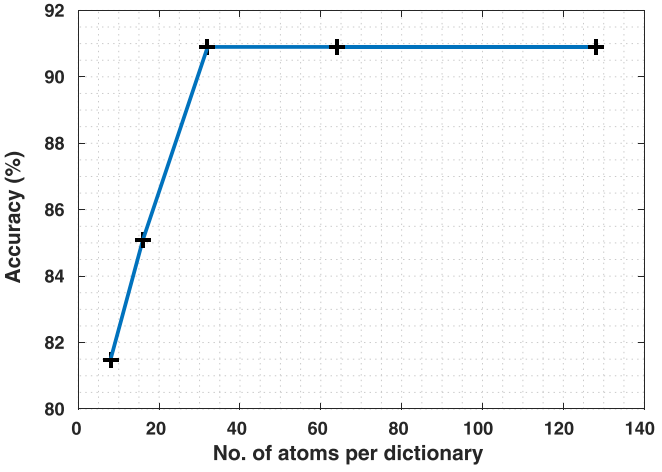

FIG. 9. (Color online) Number of chosen atoms vs classification accuracy.

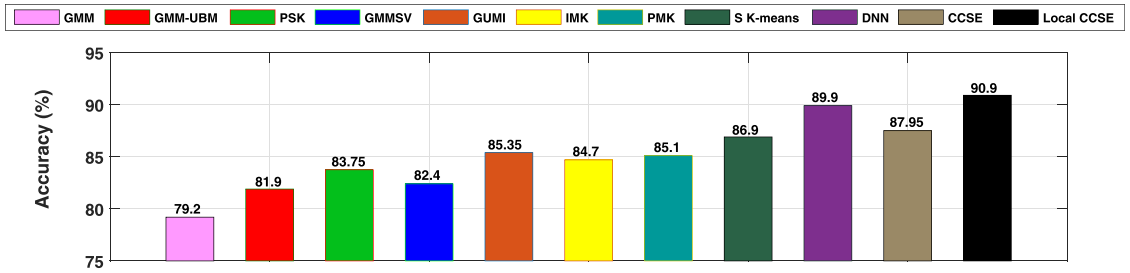

FIG. 10. (Color online) Comparison of the classification performance of the proposed local CCSE-based framework with various comparative methods.

methods is illustrated in Fig. 10. It is evident from the figure that local CCSE-based framework outperforms the other methods considered in this study. The classification accuracy obtained using the proposed local CCSE-based framework is higher than the GMM, GMM-UBM, and SVM powered by various dynamic kernels. The local CCSE-based framework shows a relative improvement of 14.77%, 10.99%, 8.54%, 10.32%, 6.45%, 7.32%, and 6.82% over classification accuracies of GMM, GMM-UBM, PSK, GMMSV, GUMI, IMK, and PMK, respectively. Also, a relative improvement of 4.6% is observed over the framework using random forest and unsupervised feature representations obtained using spherical K-means. However, the performance of DNN is comparable to the proposed framework. A small relative improvement of 1.11% is obtained by the proposed framework over the classification accuracy achieved by DNN. Also, the local CCSE outperforms CCSE by a relative improvement of 3.89%.

E. Robustness comparison

The performances of most of classification frameworks are known to degrade when training and testing conditions vary significantly. For the task in hand, these variations can arise due to difference in the recording ambiance and difference in recording devices (e.g., omni-directional vs directional microphones). We conduct an experiment to analyze the robustness of the proposed framework against differences in recording environments. Five recordings of each of the 50 species, considered in this study, are downloaded from Xeno-Canto[32] which is a crowd-sourced bird vocalization database. The recording conditions of the Xeno-Canto audio recordings (XC) are different from the recordings in the dataset used for classification comparison in previous sub-section.

XC recordings are used for testing while all the recordings used in previous experiments are used for training (75% of the vocalizations for dictionary learning and 25% for training SVM). The performance of the proposed framework and other classification methods is depicted in Fig. 11. The analysis of Fig. 11 shows that proposed local CCSE framework shows a relative improvement of 10.94%, 8.68%, 7.52%, 6.67%, 6.8%, 5.53%, 6.23%, 5.12%, 2.04%, and 3.49% over classification accuracies of GMM, GMM-UBM, PSK, GMMSV, GUMI, IMK, PMK, SK-means, DNN, and CCSE, respectively. This shows that the proposed framework is more robust to the mismatched conditions in comparison to the other comparative methods.

VI. CONCLUSION

In this work, we proposed a local CCSE-based framework for bird species identification using audio recordings. We demonstrated that local CCSE provides good species discrimination and can be used as a feature representation in a classification framework. By utilizing super-frames, information about time-frequency modulations are effectively utilized. Apart from this, we also used a restricted version of AA which only processes the data points around the boundary to find archetypes. To reduce the size of archetypal dictionaries, we proposed a greedy iterative procedure which chooses a subset of atoms from each dictionary such that the gross-correlation across atoms of all the dictionaries is decreased. Experimental evaluation showed that the local CCSE-based framework outperformed all the existing methods considered in this study. The framework also performed well when there was a difference in training-testing recording conditions.

Future work will include enforcing the group sparsity for obtaining CCSE. This can further enhance the discriminative properties of local CCSE. Also, instead of using the simple linear classifier such as linear SVM, incorporating the ensemble classifiers like random forest and neural

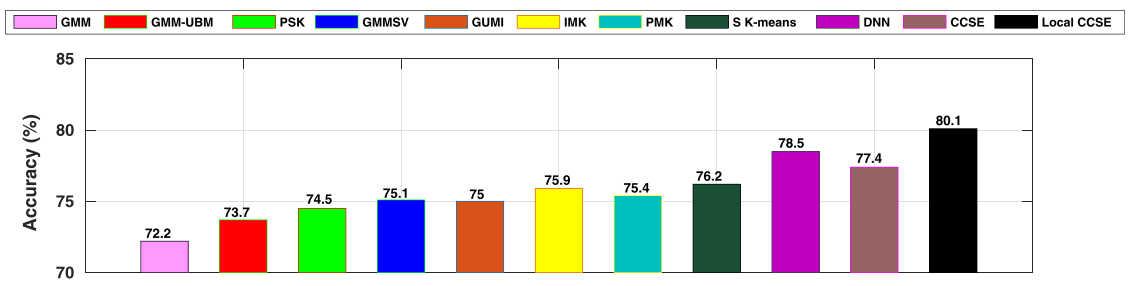

FIG. 11. (Color online) Classification performance of different methods on Xeno-Canto recordings.

networks can improve the classification performance of the local CCSE-based representation.

ACKNOWLEDGMENT

This work is partially supported by IIT Mandi under the project IITM/SG/PR/39 and Science and Engineering Research Board, Government of India under the project SERB/F/7229/2016-2017.

[1]M. Clout and J. Hay, "The importance of birds as browsers, pollinators and seed dispersers in New Zealand forests," N. Z. J. Ecol. **12**, 27–33 (1989).
[2]T. S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," Bird Conserv. Int. **18**(S1), S163–S173 (2008).
[3]C.-H. Lee, C.-C. Han, and C.-C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," IEEE/ACM Trans. Audio, Speech, Language Process. **16**(8), 1541–1550 (2008).
[4]D. E. Kroodsma, E. H. Miller, and H. Ouellet, *Acoustic Communication in Birds: Song Learning and Its Consequences* (Academic, New York, 1982), Vol. 2.
[5]A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," IEEE Trans. Signal Process. **45**(11), 2740–2748 (1997).
[6]A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proceedings of Int. Conf. Acoust. Speech, Signal Process* (May 2004), pp. 701–704.
[7]P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," IEEE/ACM Trans. Audio, Speech, Lang. Process. **14**(6), 2252–2263 (2006).
[8]P. Somervuo and A. Harma, "Bird song recognition based on syllable pair histograms," in *Proceedings of Int. Conf. Acoust. Speech, Signal Process.* (May 2004), Vol. 5, pp. V–825.
[9]S. Fagerlund, "Bird species recognition using support vector machines," EURASIP J. Appl. Signal Process. **2007**(1), 038637.
[10]D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," PeerJ **2**, e488 (2014).
[11]E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," in *CLEF (Working Notes)* (2016), pp. 547–559.
[12]B. P. Tóth and B. Czeba, "Convolutional neural networks for large-scale bird song classification in noisy environment," in *CLEF (Working Notes)* (2016), pp. 560–568.
[13]K. J. Piczak, "Recognizing bird species in audio recordings using deep convolutional neural networks," in *CLEF (Working Notes)* (2016), pp. 534–543.
[14]R. Narasimhan, X. Z. Fern, and R. Raich, "Simultaneous segmentation and classification of bird song using CNN," in *Proceedings of Int. Conf. Acoust. Speech, Signal Process.* (March 2017), pp. 146–150.
[15]D. Chakraborty, P. Mukker, P. Rajan, and A. Dileep, "Bird call identification using dynamic kernel based support vector machines and deep neural networks," in *Proceedings of Int. Conf. Mach. Learn. App.* (December 2016), pp. 280–285.
[16]A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," IEEE Trans. Neural Net. Learn. Syst. **25**(8), 1421–1432 (2014).
[17]V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proceedings of Int. Conf. Acoust. Speech, Signal Process.* (March 2016), pp. 6445–6449.
[18]P. Giannoulis, G. Potamianos, P. Maragos, and A. Katsamanis, "Improved dictionary selection and detection schemes in sparse-CNMF-based overlapping acoustic event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)* (2016), pp. 25–29.
[19]N.-C. Wang, R. E. Hudson, L. N. Tan, C. E. Taylor, A. Alwan, and R. Yao, "Change point detection methodology used for segmenting bird songs," in *Proceedings of Int. Conf. Signal Info. Process.* (2013), pp. 206–209.
[20]J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," IEEE Trans. Info. Theory **52**(9), 4036–4048 (2006).
[21]P. Frankl and H. Maehara, "The Johnson−Lindenstrauss lemma and the sphericity of some graphs," J. Comb. Theory, Ser. B **44**(3), 355–362 (1988).
[22]I. Tosic and P. Frossard, "Dictionary learning," IEEE Signal Process. Mag. **28**(2), 27–38 (2011).
[23]Y. Chen, J. Mairal, and Z. Harchaoui, "Fast and robust archetypal analysis for representation learning," in *Proceedings of Comp. Vis. Pattern Recog.* (June 2014), pp. 1478–1485.
[24]V. Abrol, P. Sharma, and A. K. Sao, "Identifying archetypes by exploiting sparsity of convex representations," in *Workshop on The Signal Processing with Adaptive Sparse Structured Representations (SPARS)* (2017).
[25]Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Proceedings of Comp. Vis. Pattern Recog.* (June 2011), pp. 1697–1704.
[26]S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," Random Struct. Algorithms **22**(1), 60–65 (2003).
[27]A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Rényi entropy based mutual information for semi-supervised bird vocalization segmentation," in *Proceedings of MLSP* (September 2017).
[28]S. Mair, A. Boubekki, and U. Brefeld, "Frame-based data factorizations," in *Proceedings of Int. Conf. Mach. Learn.* (August 2017), Vol. 70, pp. 2305–2313.
[29]L. Maaten and G. Hinton, "Visualizing data using t-sne," J. Mach. Learn. Res. **9**(Nov.), 2579–2605 (2008).
[30]"Art-sci center, University of California," http://artsci.ucla.edu/birds/database.html/ (Last viewed October 10, 2017).
[31]"Macaulay library," http://www.macaulaylibrary.org/ (Last viewed November 14, 2017).
[32]"Xeno-canto," http://www.xeno-canto.org (Last viewed October 14, 2017).